Indonesian Journal of Science & Technology

Journal homepage: http://ejournal.upi.edu/index.php/ijost/

IJoST Indonesian Journal of Science & Technology

Question Generator System of Sentence Completion in TOEFL Using NLP and K-Nearest Neighbor

Lala Septem Riza[1*], Anita Dyah Pertiwi[1], Eka Fitrajaya Rahman[1], Munir[1], Cep Ubad Abdullah[2]

[1] Department of Computer Science Education, Universitas Pendidikan Indonesia, Bandung, Indonesia
[2] Fakultas Pendidikan Ilmu Pengetahuan Sosial, Universitas Pendidikan Indonesia, Bandung, Indonesia
Correspondence: E-mail: lala.s.riza@upi.edu

ABSTRACT

Test of English as a Foreign Language (TOEFL) is one of learning evaluation forms that requires excellent quality of questions. Preparing TOEFL questions using a conventional way certainly spends a lot of time. Computer technology can be used to solve the problem . Therefore , this research was conducted in order to solve the problem of making TOEFL questions with sentence completion type. The built system consists of several stages: (1) input data collection from foreign media news sites with excellent English grammar quality; (2) preprocessing with Natural Language Processing (NLP); (3) Part of Speech (POS) tagging; (4) question feature extraction; (5) separation and selection of news sentences; (6) determination and value collection of seven features; (7) conversion of categorical data value (8) target classification of blank position word with K-Nearest Neighbor (KNN); (9) heuristic determination of rules from human experts ; and (10) options selection or distraction based on heuristic rules. After conducting the experiment on 10 news, it is obtained that 20 questions based on the results of the evaluation showed that the generated questions had very good quality with percentage of 81.93% (after the assessment by the human expert), and 70% wthe same blank position from the historical data of TOEFL questions. So, it can be concluded that the generated question has the following characteristics: the quality of the result follows the data training from the historical TOEFL questions, and the quality of the distraction is very good because it is derived from the heuristics of human experts.

ARTICLE INFO

Article History:
Submitted/Received 25 Mar 2019
First revised 25 May 2019
Accepted 05 Jul 2019
First available online 07 Jul 2019
Publication date 01 Sep 2019

Keywords:
TOEFL,
Automatic question generation,
Natural Language Processing,
Machine Learning,
K-Nearest Neighbor,
Education,
Learning

1. INTRODUCTION

Educational evaluation is a process of describing, obtaining, and presenting useful information for assessing decision alternatives in learning (Stufflebeam, 1971). One of the forms of evaluation that requires quality questions is a Test of English as a Foreign Language (TOEFL). It is the most well-known test in the field of ELT (English Language Teaching) (Alderson and Hamp-Lyons, 1996). There are several types of TOEFL, which named as (i) listening comprehension ; (ii) structure - and - written expression, and (iii) reading comprehension. In the structure and written expression in TOEFL, the parameter to be tested relates to the understanding grammar in sentences . Two types of questions in the section have been known in the TOEFL test: completion and error detection . In the first type, it is only to fill in the blank questions in TOEFL, while the second test is to choose an underline word, which is the incorrect word in the sentence . Then, because of the need of updating TOEFL questions on a regular basis with the latest topics and many questions . This makes the TOEFL questions to be helpful automatically in the process of producing qualified questions, especially on the sentence completion type. By using the existing techniques in Machine Learning, the quality of the generated questions can be kept, in accordance with the standards in the previous TOEFL questions. Nilsson (1998) explained that machine learning is a field of science to make a machine or computer to be smart . Machine learning is the most important to make the process simpler. From the machine learning methods , there is one of the most well -known algorithms , namely KNN (a machine learning algorithm for system classification). Then, as well-known, one of the techniques of processing data (namely NLP) can help the techniques to perform text processing . It is a research and application area that explores how computers can be used to understand and manipulate text (Chowdhury, 2003).

This study was focused on generating sentence completion types, in which they were generated from news articles using a combination of some following techniques , such as NLP, KNN, and heuristic techniques. The proposed system involving these methods compute articles that have good English grammar as input data ; then it produces some chosen sentence-completion questions with the answers.

Some related works can be found in the literature. For example, research conducted by Aldabe *et al.* (2006) introduced ArikIturri, which is an application used for generating fill -in -the -blank questions using NLP combined by Corpora considering the morphology and syntaxes aspects. Text2Test proposed by Aquino *et al*. (2011) utilized : text processing , scoring , and question over generation to build questions . Araki *et al*. (2016) generated multiple choice -typed questions in the subject of biology. Questions are generated by using the question template in the wh -question format . A learning management system embedded by the examination paper generated automatically was proposed by Cen *et al*. (2010). A technique with its evaluation for generating multiple choice close questions in English grammar and vocabulary (Goto *et al.,* 2010).

2. MATERIALS AND METHODS

2.1. The Method For Generating Sentence Completion Typed Questions

As shown in **Figure 1**, the computational model for generating questions can be divided into two processes: learning step and testing step that involve different data sets (*i.e.*, data training and data testing). The first is used to generate templates of questions from historical TOEFL data as data training , while the second one uses news or articles as a data testing as a candidate question.

But , both stages consist of the same following processes : inputting data , pre-processing with regex , tokenization , POS tagging with Stanford Core NLP, calculating values according to defined features , and converting categorical into numerical values. After that , results from both stages are inputted into KNN for determining a word position as the blank . Some heuristics are defined to select reasonable dummy answers for a distraction . After completing these processes , we obtained full questions with optional answers . Additionally , we can explain these processes in detail in the following section.

2.2. Data Gathering for Training and Testing

There are two sets of data required in this system, which are for training and testing. Data training is taken from historical data of TOEFL question as a reference in determining the blank position on a question. Data training is required so that the generated-question quality can be maintained as to the quality of previous TOEFL questions. For example, in this research experiment we used data training taken from the book as follows:

1. *TOEFL Exam Success from Learning Express* (Chesla, 2002).
2. *The Official Guide to the TOEFL Test Fourth Edition* (ETS, 2003).
3. *Peterson's Master TOEFL Vocabulary* (Davy and Davy, 2006).
4. *Longman Complete Course for the TOEFL Test: Preparation for the Computer and Paper Tests* (Phillips, 2001).
5. *TOEFL Practical Strategy for The Best Scores* (Pardiyono, 2005).
6. *Easy TOEIC: Test of English for International Communication* (Riyanto, 2011a).
7. *Easy TOEFL* (Riyanto, 2011b).

On the other hand, data testing, taken from news articles on internet sites that are believed to have good grammar quality, is required to be used as a question candidate. In other words, all the sentences in the selected news site can be question candidates that will be generated. For example, in this research experiment, we picked some articles from the following news: (i) Ars Technica (http://arstechnica.com), (ii) BBC News (http://bbc.com), (iii) Bloomberg (http://bloomberg.com), (iv) NBA (http://global.nba.com), (v) Forbes (http://forbes.com), (vi) People (http://people.com), (vii) Reuters (http://reuters.com) (viii) The Guardian (http://theguardian.com), (ix) The Star (http://thestar.com), and (x) VOA News (http://voanews.com).

2.3. Data Processing

Pre-process was done on two types of datasets (i.e., data training and data testing). The first stage of removal of punctuation (regex). The removed-punctuations were other than dots and underscores. A dot is used for a marker or separator between sentences. Whereas, underscore is used to mark the blank position during feature extraction. Then, the other pre-process stage is *tokenization* which is the stage to divide one sentence into a word. This stage is necessary to simplify the process *of part-of-speech tagging* and feature extraction in the next stage. For example, given a complete sentence as follow: *One of the most popular Indonesian products is Batik, it has been internationally recognized.* So, after these processes, we obtain the following sequence: " One | of | the | most | popular| Indonesian |product | is | Batik | it | has | been| internationally | recognized ". Thus, it can be seen that the complete sentence separated into word by word . After that , these processes are also applied to all data.

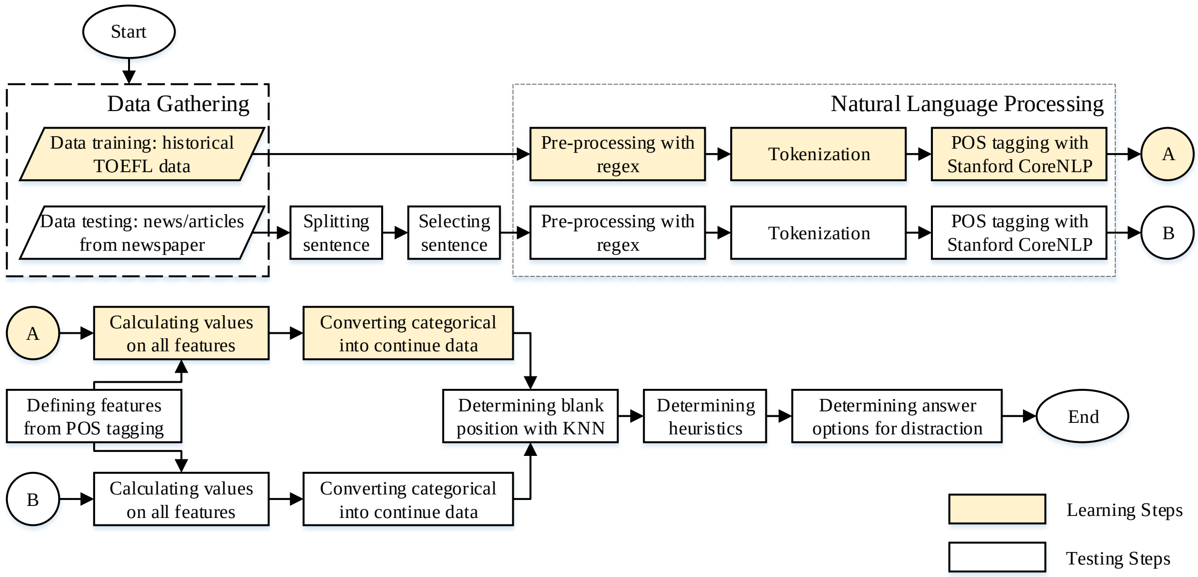
Start
Data Gathering
Data training: historical TOEFL data
Data testing: news/articles from newspaper
Splitting sentence
Selecting sentence
Natural Language Processing
Pre-processing with regex
Tokenization
POS tagging with Stanford CoreNLP
A
Pre-processing with regex
Tokenization
POS tagging with Stanford CoreNLP
B
A
Calculating values on all features
Converting categorical into continue data
Defining features from POS tagging
Determining blank position with KNN
Determining heuristics
Determining answer options for distraction
End
B
Calculating values on all features
Converting categorical into continue data
Learning Steps
Testing Steps

Figure 1. Flow model of question generator system.

2.4. Part of Speech (POS) Tagging by Stanford Core NLP

At this stage, every word will come to part-of-speech tagging (POS Tagging) process to get information about the word class which will then be needed for feature extraction. There were many computational linguistic software. One of them used in this study is Stanford Core NLP which can be accessed on page https://stanfordnlp.github.io/CoreNLP/. It is a toolkit created for research purposes in the NLP field (Manning , *et al.,* 2014). There are 8 English word classes, namely noun, verb, pronoun, preposition, adverb, conjunction, adjective, and articles. Moreover, there is a popular and commonly used tag set, which is Penn Treebank Tag set (Marcus *et al.,* 1993). The examples of the use of POS Tagging in the previous sentence on the preprocessing data is "CD|IN|DT|RBS|JJ|JJ| NNS|VBZ|NNP|PRP| VBZ|VBN|RB|VBN", where CD means cardinal number. IN, DT, RBS, JJ, NNS, VBZ, NNP, PRP, VBN, and RB are the meaning in preposition, determiner, adverb, adjective, noun plural, present verb for the 3rd person, proper noun, personal pronoun, verb past participle, and adverb, respectively . Thus, each word has its own part-of-speech label. This will facilitate the process of generating questions in the next stages.

2.5. Separation and Selection of Sentences from News Articles

Basically, there are two steps in this section, as follows: separation and selection. The first one is the process of separating sentences in one long news text. It is done to make it easier in making a question since usually it only contains one sentence. So, it is that one large text contains thousands line of sentences will be separated from dot (.) to dot (.). The way of separation of this sentence is by using the regex command. It will search for the mentioned punctuation and then use the split function on any punctuation that has been found.

At the selection stage, the selection of sentences is done to simplify and shorten the classification process. Selection of sentences was done with considering two conditions, as follows:

1. Sentences consist of 10 to 30 words. This requirement has been discussed with the previous expert.

2. Then sentences are randomly chosen from the first condition.

Based on the two requirements above, the sentences in the news which became the question candidate is expected to be more qualified.

2.6. Determination of Seven Features on Data Sets

This stage is the process of determining the feature to be used as a word attribute for the blank position classification. These features are important ones that facilitate the classification process. This feature consists of seven features as proposed by this study as follows (Hoshino and Nakagawa, 2005):

- *Post*: It is a column filled by the Tag POS of the word in column 1 of that line. The post column is auto-filled using the Stanford CoreNLP library.
- *Prev_Pos*: It is a column containing the POS Tag of the previous word in one sentence. This is not different from other post columns. This column is auto-filled using the Stanford CoreNLP library.
- *Next_Pos*: It is a column containing the POS Tag of the next word in one sentence. This column is automatically filled using the Stanford CoreNLP library.
- *Position*: It is a column filled with a number which is the position of the word in that line in one sentence. For example, if in a sentence there are 10 words, then the order of words is 1-10, the column position in the first word will be filled by the number 1.
- *Sentence*: It is a column containing numbers to determine the number of words in a single sentence. If you have a sentence containing 10 words, this sentence column will contain the number 10 from the first word line to the last line of words.
- *Word-Length*: It is filled by the number of words that are repeated in one sentence. For example, when in a sentence there are 2 words of, then, this word-length column will contain number 2 in the word line.
- *Word*: It is a column containing words in sentences that have passed through the tokenization process in the system.

Moreover, *Target*: It is an output feature showing the index of the blank position.

2.7. Value Calculation of Seven Features on Data Sets

It is the process of collecting seven features for the set of data. In collecting these seven features, it took advantage of a process that has been done before. POS features, next word POS, and previous word POS are taken from tokenization and POS Tagging. In addition, the position feature uses the function count on each sentence to know the order of the words position in the sentence. The function of count is used to calculate how many words in a single sentence and then the result will be the value for sentence feature. Furthermore, the word length feature also utilizes the count function, but before that, the words in one sentence must be compared, so the word length of a word will increase if the word is repeated several times in one sentence. Meanwhile, the target is a determinant of word classification and features. The target of data training is automatically obtained by the system by detecting whether there is an underscore (_) at the end of the word. If there is an underscore, then the word is a blank position in the sentence, meaning the target is true.

Table 1. Example of data training of seven features values.

Words	*POS*	*Prev_POS*	*Next_POS*	*Position*	*Sentence*	*Word-Length*	*Target*
One	1	CD	0	IN	1	14	FALSE
Of	1	IN	CD	DT	2	14	FALSE
Of	1	IN	CD	DT	2	14	FALSE
the	1	DT	IN	RBS	3	14	FALSE
Most	1	RBS	DT	JJ	4	14	FALSE
Popular	1	JJ	RBS	JJ	5	14	FALSE
Indonesian	1	JJ	JJ	NNS	6	14	FALSE
Products	1	NNS	JJ	VBZ	7	14	FALSE
Is	1	VBZ	NNS	NNP	8	14	FALSE
Batik	1	NNP	VBZ	PRP	9	14	FALSE
It	1	PRP	NNP	VBZ	10	14	FALSE
has	1	VBZ	PRP	VBN	11	14	FALSE
been	1	VBN	VBZ	RB	12	14	FALSE
internationally	1	RB	VBN	VBN	13	14	TRUE
recognized	1	VBN	RB	0	14	14	FALSE

An example of value collecting of these seven features is in **Table 1** . As the example, of the following sentence “one of the most popular indonesian batik , it has been internationally recognized ” from data training.

2.8. Converting Categorical Data to Continuous

KNN is a distance calculation algorithm that necessarily requires numerical data to find the closest distance to be able to determine the target. So, the point we need in order to find the distance is to convert the categorical data into numerical data. The categorical data of the 7 features in this research are part -of -speech and word . These categorical data will be converted into continuous data using the following equation:

$$S_x = \frac{100}{(P_{x\,max} - P_{x\,\mathrm{min})}} \qquad (1)$$

$$O_x = -P_{x\,min} \qquad (2)$$

$$V = (S_x\, x\, P_x + O_x, \ldots, S_n\, x\, P_n + O_n \qquad (3)$$

where:

- *S* is the calculation of categorical data initialization data
- 100 is the range that can be changed and determined as needed
- *P* is the class of each categorical data , which is *P1* = part of speech , *P2* = previous word of part of speech , *P3* = next word of part of speech.
- *x* is the index of categorical data classes of *n*
- *V* is the categorical data vector after it is converted to numeric (continuous)

POS tags are initialized into numbers for easy calculation . The initialization is based on proximity between tags . The closer the tag, the tag has the proximity or Similarity as shown in **Table 2**.

Table 2. Initialization values of POS tags.

1 = CC	2 = CD	3 = DT	4 = EX	5 = FW	6 = IN	7 = JJ	8 = JJR
9 = JJS	10 = LS	11 = MD	12 = NN	13 = NNS	14 = NNP	15 = NNPS	16 = PDT
17 = POS	18 = PRP	19 = PRP$	20 = RB	21 = RBR	22 = RBS	23 = RP	24 = SYM
25 = TO	26 = UH	27 = VB	28 = VBD	29 = VBG	30 = VBN	31 = VBP	32 = VBz
33 = WDT	34 = WP	35 = WP$	36 = WRB				

Since the three categorical data have the same data content, the calculation is as follows:

$$S_1 = \frac{100}{36-1} = 2{,}86, \quad O_1 = -1$$
$$S_2 = \frac{100}{36-1} = 2{,}86, \quad O_2 = -1$$
$$S_3 = \frac{100}{36-1} = 2{,}86, \quad O_3 = -1$$

The next step is to calculate V.

$$V = (S_1\, x\, P_1 +\ O_1, S_2\, x\, P_2 +\ O_2, S_3\, x\, P_3 +\ O_3)$$
$$= (2{,}86\, x\, P_1 - 1,\ 2{,}86\, x\, P_2 - 1,\ 2{,}86\, x\, P_3 - 1)$$

By applying the same questions as above, we calculate data with features as showed in **Table 1**. For example, the conversion of the word 'one' is as follows:

$$V = (2{,}86\, x\, P_1 - 1, 2{,}86\, x\, P_2 - 1, 2{,}86\, x\, P_3 - 1)$$
$$= (2{,}86\, x\, CD - 1, 0, 2{,}86\, x\, IN - 1)$$
$$= (2{,}86\, x\, 2 - 1, 0, 2{,}86\, x\, 6 - 1)$$
$$= (4{,}72, 0, 16{,}16)$$

After the calculation is obtained, perform the calculation to all words in one sentence. Thus, the results will be obtained as showed in **Table 3**. This calculation applied to all data, both data training and data testing.

Table 3. The example of value conversion of **Table 1**.

Words	POS	Prev_POS	Next_POS	Position	Sentence	Word-Length	Target
One	4.72	0.00	16.16	1	14	1	FALSE
Of	16.16	4.72	7.58	2	14	1	FALSE
Of	7.58	16.16	61.92	3	14	1	FALSE
the	61.92	7.58	19.02	4	14	1	FALSE
Most	19.02	61.92	19.02	5	14	1	FALSE
Popular	19.02	19.02	36.18	6	14	1	FALSE
Indonesian	36.18	19.02	90.52	7	14	1	FALSE
Products	90.52	36.18	39.04	8	14	1	FALSE
Is	39.04	90.52	50.48	9	14	1	FALSE
Batik	50.48	39.04	90.52	10	14	1	FALSE
It	90.52	50.48	84.8	11	14	1	FALSE
has	84.80	90.52	56.20	12	14	1	FALSE
been	56.20	84.80	84.80	13	14	1	TRUE
internationally	84.80	56.20	0.00	14	14	1	FALSE
recognized	4.72	0.00	16.16	15	14	1	FALSE

Table 4. Example of seven features of value in data testing

Words	POS	Prev_POS	Next_POS	Position	Sentence	Word-Length
Industry	33.32	0	32.32	1	11	1
Aggregate	32.32	33.32	36.18	2	11	1
Averages	36.18	32.32	87.66	3	11	1
Are	87.66	36.18	84.80	4	11	1
Calculated	84.80	87.66	81.94	5	11	1
Starting	81.94	84.8	16.16	6	11	1
With	16.16	81.94	7.58	7	11	1
A	7.58	16.16	19.02	8	11	1
market-cap	19.02	7.58	19.02	9	11	1
Weighted	19.02	19.02	33.32	10	11	1
Average	33.32	19.02	0.00	11	11	1

2.9. Determination of Blank Position with KNN

This stage is a step done to determine the target of each word in the data testing. This targeting is done by comparing word by word in the data testing of sentence with the word on data training. The word along with the seven features will be calculated the distance to determine the target classification of words that will be the blank position in the sentence. The distance calculation at KNN stage uses Euclidean Distance formula. So, the distance between words from the data training and the data testing can be calculated. For example, given in **Table 4** to be data testing that will be calculated the distance to data training.

From **Tables 3** and **4**, we examine the Euclidean Distance . So, the word distance counted is the word 'One' and the word ' Industry':

$$d = \sqrt{\begin{matrix}(33{,}32 - 4{,}72)^2 + (0 - 0)^2 + \ (32{,}12 - 16{,}16)^2 + \\ (1 - 1)^2 + (11 - 14)^2 \ + (1 - 1)^2 + 1^2\end{matrix}} = 33{,}99$$

The calculation is performed to all data testing compared with data training. Thus, it will be obtained *k* the nearest distance on each word. As the example, *k* is 3. Thus, the most target of three closest distances for a word will be the target of the data testing. If two of the targets are false, then the target data testing will be false as well. From that target, the position of blank can be determined. True target to the word in one sentence will be a blank position.

2.10. Determination of Heuristics and Distractors

This stage is the process of heuristic determination to produce a distraction. The heuristic is structured for the purpose of producing qualified distractions. The rules for generating distractions are as follows: verb, preposition, pronoun, modal, determiner, conjunction, *wh*-pronoun, *wh*-determiner, and *wh*-possessive, and *wh*-adverb. For example, verbs have several types of tags, namely VB, VBD, VBG, VBN, VBP, and VBZ. Selecting distraction in a verb is taken from an online English dictionary using the Application Programming Interface (API) of Ultra lingua that can be accessed at http://api.ultralingua.com/. This API will generate all possible equivalents of a similar word from the verb with a feature called 'verb conjugation'.

After determining heuristics, the last stage is to generate incorrect answers to be distraction. Since a question with the type of sentence completion has four options with one correct answers and three wrong answers, we need to choose three distractions. Therefore, if the POS tagging of true answer is VBZ, then the distractions could be the corresponding verb with the POS tagging VBD, MD VB, and VB. For example in a question:

The earth spins on its axis and ... 23 hours, 56 minutes and 4.09 seconds for one complete rotation.

A. Needed
B. Will need
C. Need
D. Needs

3. RESULTS AND DISCUSSION

3.1. Experimental Design

At the experimental stage, the system will implement the model that has been created and produce the TOEFL questions with the type of sentence completion question. The numbers of generated question are 50 questions consisting of 1% of the data training which is 30 questions and 20 questions from 10 different news articles. As mentioned earlier, there are 10 news websites with different topics as data testing, which is listed in **Table 5**.

After conducting the experiments, there are three kinds of analyses that will be done, as follows:

1. Same blank position analysis: This analysis will prove the accuracy of the system in choosing the blank position. Generated-question from the data training will be calculated how many blank positions which are the same in order to get the level of accuracy.
2. Consistency of the answers analysis: This analysis is an analysis that involving some experts to answer the generated-questions. Two experts will answer the following questions and they will be checked whether the answers filled by the experts have the same answers with the provided answer key.
3. Evaluation and analysis on the quality of questions from the expert: The generated-questions will be evaluated by two experts in order to determine its quality. The assessment given by the experts is based on four metrics of assessment proposed by Araki et al (2016) as follows:

   a) Grammatical Correctness (GC): It determines whether a question is syntactically well formed. The researchers determine three points to show the scale of the matrix based on the number of grammar error. Grammar error is calculated except for distraction that makes the sentence wrong. The values of the scale are 1, 2, and 3 that mean the question has no grammatical errors (best), the question has 1 or 2 grammatical errors, and the question has 3 or more grammatical errors, respectively.

   b) Answer Existence (AE): It identifies whether the answer to a question can be inferred from the related part of the question. The researchers determine two points as follows: 1 (yes) means the answer to a question can be inferred from the question, and 2 (no) means the answer to a question cannot be inferred from the question.

c) Distractor Quality (DQ): It is an assessment to measure how precisely a distractor from the four underline is raised. The researcher made a two-point scale for this assessment as follows: 1 (worst) means Distractor can be easily identified as wrong answers, 2 (best) means Distractor can be feasible.

d) Difficulty Index (DI): It is an assessment of how difficult the generated question from the system. This assessment is determined from the overall aspects of both questions and distractions. The researchers made a scale of three points as follows: 1 (easy) means the generated question is considered easy, 2 (medium) means the generated question is considered sufficient, and 3 (hard) mean the generated question is considered very difficult.

3.2. Experimental Results

After executing the proposed system as explained in the previous section, we obtained 50 generated questions. **Table 6** contains of some questions that have been generated. It can be seen that the yellow color on the word is the correct answer.

Table 5. News sites data used in the study.

No	File Names	News' URL	Titles	Topics
1	arc.txt	https://arstechnica.com/gaming/2018/04/more-human-than-human-review-light-on-killer-robots-killer-on-ai-inspection/	As AI advances rapidly, More Human Than Human says, "Stop, let's think about this"	Technology
2	bbc.txt	http://www.bbc.com/news/world-asia-44100278	Surabaya church attacks: One family responsible, police say	News
3	bloomberg.txt	https://www.bloomberg.com/news/articles/2018-03-29/shadow-cast-over-peace-talks-as-fighting-flares-in-south-sudan	Cast over peace talks as fighting flares in South Sudan	Politic
4	nba.txt	http://global.nba.com/news/drummond-has-another-big-performance-pistons-beat-knicks-115-109/?cid=trafficdriver:nbacom:homepage	Drummond Has Another Big Performance to push Pistons past Knicks	Sport
5	forbes.txt	https://www.forbes.com/sites/davidthier/2018/03/31/3-reasons-why-fortnites-comet-could-actually-destroy-tilted-towers-and-when/#7e3bacb423c3	3 Reasons Why Fotnites Comet Could Actually Destroy Tilted Towers	Technology

Table 5 (continue). News sites data used in the study.

No	File Names	News' URL	Titles	Topics
6	people.txt	http://people.com/food/wedding-registry-items-worth-asking-for/	Wedding Registry Items Worth Asking For	Food
7	reuters.txt	https://www.reuters.com/article/us-citigroup-results/citigroup-profit-beats-on-strength-in-consumer-banking-equity-trading-idUSKBN1HK1MR?il=0	Citigroup profit beats on strength in consumer banking, equity trading	Business
8	theguardian.txt	https://www.theguardian.com/lifeandstyle/2018/mar/18/reading-rooms-the-story-of-an-authors-house	Reading rooms: the story of an author's house	Life Style
9	thestar.txt	https://www.thestar.com.my/news/education/2018/04/29/education-is-a-right-for-all/	Education is a right for all	Education
10	voa.txt	https://www.voanews.com/a/nasa-insight-mission-to-mars/4380857.html	Nasa Insight Mission to Mars	Science

3.3. Discussion

In this section, it will be explained the analysis of the results obtained based on the experimental design in the previous section, namely same blank position analysis, consistency of the answer analysis, and evaluation and quality analysis of the questions by the experts.

3.3.1. The Same Blank Position Analysis

As the previous explanation, this analysis will prove the accuracy of the blank position between the generated sentences from the data training with the original question.

Table 7 shows the value of the same blank position. The number 1 indicated that the blank position on the generated -question by the system was equal to the blank position on the original question . There are 21 questions from 30 questions with 70 % has the same blank position . These results indicated that there were still generated -questions of the system with different blank positions . The difference may be caused by the smaller distance in the selected tag with the majority of true targets . Thus , the obtained blank position was not the same as the original question.

Table 6. Results: 50 questions and answers generated by the proposed system.

Questions and Answers

1. Some hangers, buildings used to ... large aircraft, are very tall that rain occasionally falls from clouds that form along the ceilings
    A. hold
    B. will hold
    C. held
    D. holds
2. Dairy farming is ... leading agricultural activity in the United States
    A. an
    B. a
    C. one
    D. the
3. ... this the Most Valuable Car in the World? With the death of Florida flea-market magnate Preston Henn, a vintage Ferrari is poised to test the $100 million mark.
    A. were
    B. being
    C. is
    D. will be

....

50. Weathering is the action ... surface rock is disintegrated or decomposed
    A. whereby
    B. where
    C. why
    D. when

3.3.2. Consistency of the Answer Analysis

As mentioned in the previous explanation, this analysis will prove whether the experts answered the same questions according to the answer key generated by the system. The experts will answer 30 questions from data training, and 20 questions from data testing.

In **Table 8**, the two experts answered the number of questions from the data training correctly with different amounts. Expert 1 symbolized by E1 answered 25 questions out of 30 questions correctly. The percentage is 83%. Meanwhile, the expert 2 symbolized by E2 answered 21 true questions, with 70%. Whereas, as shown in **Table 9**, 20 questions generated from data testing, Expert 1 answered 19 questions according to the answer key, with a percentage of 95%. However, expert 2 only answered 16 questions according to the answer key with the value of 80%.

Based on the results of these two experts, it can be concluded that not all questions have a good quality of answer or a good distraction. This is proved by the difference of answers from two experts with the answer key. The difference can be caused by the ambiguity in the sentence, or the existence of a distraction so that it can generate two correct answers. From this assessment, it can be drawn the average consistency of this answer is 81.25%.

Table 7. The same of blank position analysis.

ID Question	Blank Position	ID Question	Blank Position	ID Question	Blank Position
1	1	2	1	3	1
4	1	5	1	6	1
7	1	8	1	9	1
10	1	11	1	12	1
13	1	14	1	15	1
16	1	17	1	18	1
19	1	20	1	21	1
22	0	23	0	24	0
25	0	26	0	27	0
28	0	29	0	30	0

Table 8. The analysis by experts on data training (fitting step).

ID Question	Answer Key	Expert 1	Expert 2	ID Question	Answer Key	Expert 1	Expert 2
1	A	A	A	16	D	D	D
2	B	B	B	17	C	D	A
3	A	C	A	18	C	C	A
4	A	A	A	19	C	C	B
5	D	D	D	20	B	B	B
6	D	D	D	21	B	B	B
7	A	A	A	22	C	C	A
8	C	C	C	23	B	B	A
9	C	C	A	24	C	C	C
10	D	D	D	25	C	C	C
11	D	D	D	26	D	D	D
12	C	C	C	27	C	C	D
13	D	D	D	28	B	A	B
14	D	D	D	29	B	D	B
15	A	A	D	30	A	B	B

3.3.3. Question evaluation and quality analysis by Human Experts

Based on the experimental design, the questions were evaluated for quality with 4 assessment metrics: grammatical correctness (GC), answer existence (AE), distractor quality (DQ), and difficulty index (DI). The results of the evaluation wereassessed by the experts with the average assessment index of grammatical correctness 1.05, answer existence 1.09, distractor quality 1. 71, and difficulty index 1.66. Then, it will be categorized that the quality of this question with five categories, which can be classified such as very good (between 80 and 100%), good (between 60 and 80 %), enough (between 40 and 60%), less (between 20 and 40%), and very less (less than 20 %). The results of these calculations are presented in **Table 10**.

Table 9. The analysis by experts on data testing (testing step).

ID Question	Answer Key	Expert 1	Expert 2	ID Question	Answer Key	Expert 1	Expert 2
1	B	B	D	11	D	D	D
2	C	C	C	12	C	C	C
3	A	A	A	13	D	D	D
4	D	B	B	14	A	A	A
5	C	C	C	15	B	B	B
6	C	C	C	16	A	A	A
7	C	C	C	17	C	C	C
8	B	B	D	18	D	D	D
9	C	C	D	19	D	D	D
10	C	C	C	20	A	A	A

3.3.4. The Comparison with Previous Research

In this section, it will be compared the model and implementation of this study with the previous studies that have similar types of research. There were many studies related to this question generator. Some of them become the references for this study in developing the system model. Both the reference of the algorithm, the problem attributes, and the evaluation of the question quality. The comparison is shown in **Table 11**.

Table 10. The calculation results of each parameter.

Parameters	Ideal Value	Σ Score per Parameter	Percentage	Category	Explanation
Grammatical Correctness (GC)	1	1.05	95.2%	Very good	Low score shows better value.
Answer Existence (AE)	1	1.09	91.7%	Very good	Low score shows better value.
Distractor Quality (DQ)	2	1.71	85.5%	Very good	High score shows better value.
Difficulty Index (DI)	3	1.66	55.3%	enough	High score shows better value.
AVERAGE			81.93%		

Table 11. Comparison with other system.

References	Methology	Types of Questions	Language	Evaluation/Analysis Strategies
(Goto *et al.*, 2010)	Sentence extraction, determining blank position using Conditional Random Field	Multiple-choice cloze question	English	Involving expert judgment with grammatical correctness of assessment criteria and quality of blank position.
(Susanti, *et al.*, 2015)	Word selection from article using WordNet	Vocabulary test	English	Using human expert to answer and determine whether the question given is from machine-generated or human-generated
(Majumder and Saha, 2015)	Sentence selection using novel technique and Parse Tree Matching	Multiple-choice	English	Question is assessed by 5 human evaluators
(Hill and Simha, 2016)	Determining blank position using NER, and choosing distractor using Google n-gram	Multiple-choice fill-in-the-blank	English	Using human expert as much as 67 native English-speaking volunteers to give an opinion on the given question
(Pannu *et al.*, 2018)	Sentence and blank position selection using NER	Fill-in-the-blank question	English	Using human expert with 3 assessment metrics namely validity, key quality, and sentence quality
(Chen *et al.*, 2006)	Sentence selection and generate question using NLP techniques	Error detection and fill-in-the-blank	English	Using human expert to assess the feasibility of the generated-question.

Table 11 (continue). Comparison with other system.

References	Methology	Types of Questions	Language	Evaluation/Analysis Strategies
(Agarwal *et al.,* 2011)	Sentence selection using summarization, NER	Cloze question	English	Using human expert in determining whether the generated-question is applicable or not.
(Papasalouros *et al.,* 2008)	ontology-based strategies like class based, property based, terminology-based strategies	Multiple-choice	English	Using 3 assessment metrics that are pedagogical quality, linguistic correctness, and number of generated question which are then reviewed by 2 educational experts
(Huang and He, 2016)	Using semantic linguistic framework	Wh-question	English	Comparing the difficulty index of the questions generated by the system with the question generated by humans.
(Hoshino and Nakagawa, 2005)	Using KNN and Naïve Bayes in blank positioning, using 7 features	Fill-in-the-blank	English	Using the comparison between the blank positions generated by KNN and Naïve Bayes
(Araki *et al.,* 2016)	Compile the questions using available templates	Wh-question	English	Using 4 quality evaluation metrics assessment such as Grammatical Correctness, Answer Existence, Distractor Quality, and Difficulty Index
This research	NLP techniques to process sentences and KNN to specify blank positions; A heuristic to determine the divers.	Sentence completion pada TOEFL	English	Using the same blank position analysis, consistency of the answers, and 4 assessment metrics in evaluating quality of the question

4. CONCLUSION

After conducting this research, we could draw the following conclusions:

a) This research succeeded in making the computational model to produce sentence completion in TOEFL automatically using Natural Language Processing techniques, *k*-Nearest Neighbor algorithm, and heuristics. Basically, the system contains two main processes: learning and testing. Both stages consist of inputting data, pre-processing with regex, tokenization, POS tagging with Stanford CoreNLP, calculating values according to defined features, and converting categorical into numerical values. After that, results from both stages are inputted into KNN for determining a word position as the blank. Some heuristics are defined to choose reasonable dummy answers for distraction.

b) The results of the question evaluation showed that the generated-question has excellent quality with a percentage of 81.93% after analyzed by the experts, 81.25% of consistency of the answer, and 70% of the same blank position.

Based on the results and analyses, this study contributes to be used as a tool for generating questions with the sentence completion on TOEFL automatically derived from news articles.

5. AUTHORS' NOTE

The author(s) declare(s) that there is no conflict of interest regarding the publication of this article. Authors confirmed that the data and the paper are free of plagiarism.

6. REFERENCES

Agarwal, M., Shah, R., and Mannem, P. (2011). Automatic question generation using discourse cues . In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*(pp. 1-9). Association for Computational Linguistics.

Alderson, J. C., and Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback . *Language Testing*, *13*(3), 280-297.

Aldabe , I., De Lacalle , M. L., Maritxalar , M., Martinez , E., and Uria , L. (2006). Arikiturri : an automatic question generator based on corpora and nlp techniques . In *International Conference on Intelligent Tutoring Systems* (pp. 584-594). Springer, Berlin, Heidelberg.

Aquino , J. F., Chua , D. D., Kabiling , R. K., Pingco , J. N., and Sagum , R. (2011). Text 2Test : Question generator utilizing information abstraction techniques and question generation methods for narrative and declarative text . In *Proceedings of the 8th National Natural Language Processing Research Symposium* (pp. 29-34).

Araki, J., Rajagopal, D., Sankaranarayanan, S., Holm, S., Yamakawa, Y., and Mitamura, T. (2016). Generating questions and multiple -choice answers using semantic analysis of texts. In *Proceedings of COLING 2016 , the 26 th International Conference on Computational Linguistics: Technical Papers* (pp. 1125-1136).

Cen, G., Dong, Y., Gao, W., Yu, L., See, S., Wang, Q., and Jiang, H. (2010). A implementation of an automatic examination paper generation system . *Mathematical and Computer Modelling*, *51*(11-12), 1339-1342.

Chen , C. Y., Liou , H. C., and Chang , J. S. (2006). Fast : an automatic generation system for grammar tests. In *Proceedings of the COLING /ACL on Interactive presentation sessions* (pp . 1-4). Association for Computational Linguistics.

Chesla, E. (2002). TOEFL Exam success from LearningExpress . New York: LearningExpress.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51-89.

Davy , E., and Davy , K. (2006). Peterson's Master TOEFL Vocabulary . USA: Petersons Co.

ETS, TOEFL Practice TESTS Volume 1, Princeton, 2003.

Goto , T., Kojiri , T., Watanabe , T., Iwata , T., and Yamada , T. (2010). Automatic generation system of multiple -choice cloze questions and its evaluation . *Knowledge Management and E-Learning*, *2*(3), 210.

Hill , J., and Simha, R. (2016). Automatic Generation of Context-based Fill-in-the-blank Exercises using Co -occurrence Likelihoods and Google n-grams . In *Proceedings of the 11 th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 23-30).

Hoshino , A., and Nakagawa , H. (2005). A real -time multiple -choice question generation for language testing : a preliminary study . In *Proceedings of the second workshop on Building Educational Applications Using NLP* (pp. 17-20). Association for Computational Linguistics.

Huang , Y., and He , L. (2016). Automatic generation of short answer questions for reading comprehension assessment. *Natural Language Engineering*, *22*(3),457-489.

Manning , C., Surdeanu , M., Bauer , J., Finkel , J., Bethard , S., and McClosky , D. (2014). The Stanford CoreNLP natural language processing toolkit . In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60).

Majumder , M., and Saha, S. K. (2015). A system for generating multiple choice questions : With a novel approach for sentence selection . In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 64-72).

Marcus , M. P., Marcinkiewicz , M. A., and Santorini , B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313-330.

Nilsson, N. J. (1998). Introduction to Machine Learning. California, Amerika.

Pannu, S., Krishna, A., Kumari, S., Patra, R., and Saha, S. K. (2018). Automatic Generation of Fill-in- the -Blank Questions From History Books for School -Level Evaluation . In *Progress in Computing, Analytics and Networking* (pp. 461-469). Springer, Singapore.

Papasalouros , A., Kanaris , K., and Kotis , K. (2008). Automatic Generation of Multiple Choice Questions From Domain Ontologies. In *e-learning*, 427-434.

Pardiyono, (2005). TOEFL Practical Strategy for The Best Scores. Yogyakarta: ANDI.

Phillips , D. (2001). Longman Complete Course for the TOEFL Test : Preparation for the Computer and Paper Tests. New York: Pearson Education.

Riyanto, S. (2011a). Easy TOEIC: Test of English for International Communication. Yogyakarta : Pustaka Pelajar.

Riyanto, S. (2011b). Easy TOEFL. Yogyakarta: Pustaka Pelajar.

Stufflebeam , D. L. (1971). The use of experimental design in educational evaluation. *Journal of Educational Measurement* , *8*(4), 267-274.

Susanti , Y., Iida , R., and Tokunaga , T. (2015). Automatic generation of english vocabulary tests. In *CSEDU*(pp. 77-87).